Week 7 : Descriptive Statistics (Part 2)

☰ pending tasks	
☰ type	

Introduction to measures of spread - percentile

def percentile - (t = 6:49)

Learning objectives:

1. What are percentiles?
2. What are frequently used percentiles?
3. How to compute the percentile rank of a value in the data?
4. What is the effect of transformations on percentiles?
5. What are different measures of spread?
6. What is the effect of transformation on measures of spread?
7. What are box plots and how to use them to visualize some measures of centrality and spread?

Percentiles

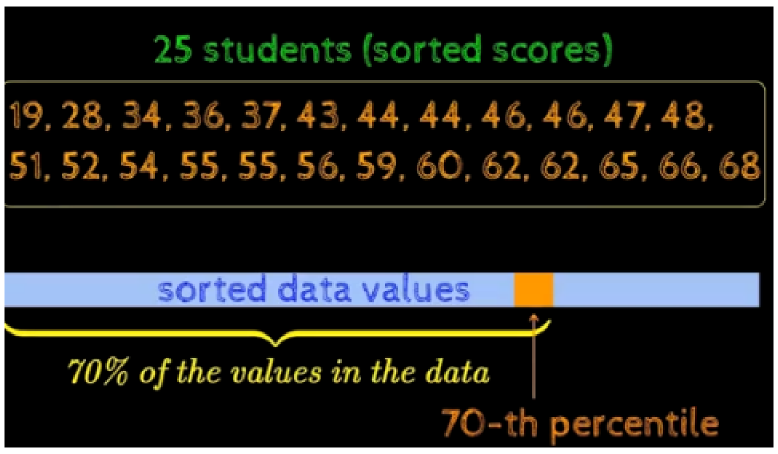


Fig.1 An example of percentile.

‘P’ percentile is the value below which ‘p’ percentage of the data falls. In the above example, the 70th percentile is the value below which 70% of the scores are found.

Example to calculate the percentile:

1. Sort the data.
2. Compute location of the p^{th} percentile.

$$L_p = \frac{p}{100}(n+1)$$

Here it is 18.2

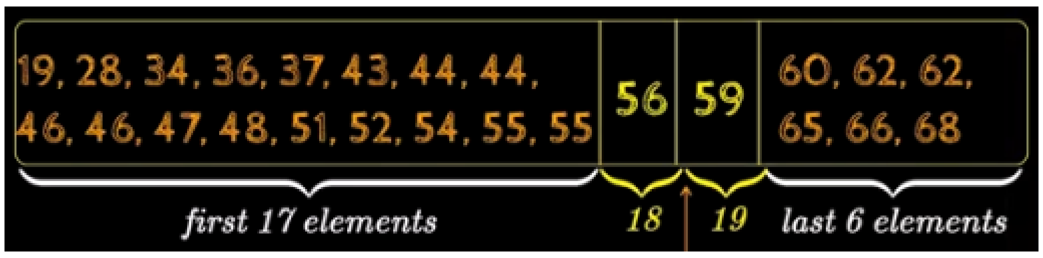


Fig.2 Location 18.2

3. Calculate the value that is (0.2) * (56 - 59), for this example. This gives the value 56.6, the 70th percentile of the given data.

Procedure to calculate percentile

1. Sort the data
2. Compute the location of the p^{th} percentile. $L_p = \frac{p}{100}(n+1)$
3. Find the integer part of $L_p = i_p$ (18 in the example above) Fractional part of $L_p = f_p$ (0.2 in the example)
4. Compute the p^{th} percentile as

$$Y_p = x_p + f_p * (x_{ip+1} - x_{ip})$$

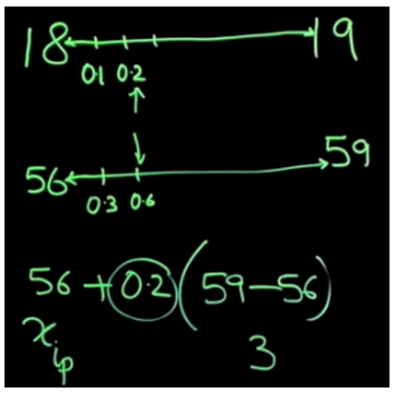
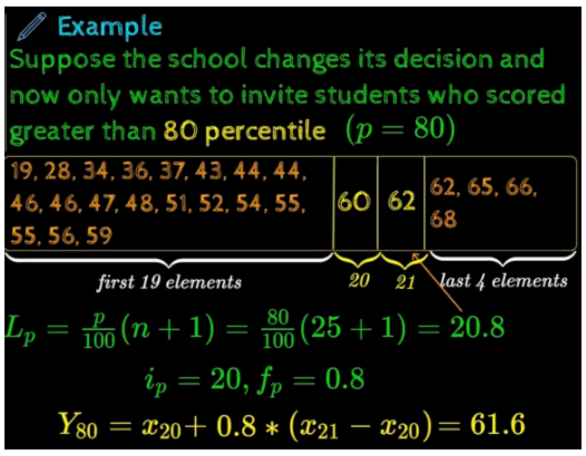


Fig.3 Intuition behind the computation.

In special cases where f_p = 0 the percentile value will be a point in the dataset.



Example

Suppose the school changes its decision and now only wants to invite students who scored greater than 80 percentile ($p = 80$)

19, 28, 34, 36, 37, 43, 44, 44, 46, 46, 47, 48, 51, 52, 54, 55, 55, 56, 59	60	62	62, 65, 66, 68
first 19 elements	*20*	*21*	*last 4 elements*

$L_p = \frac{p}{100}(n+1) = \frac{80}{100}(25+1) = 20.8$

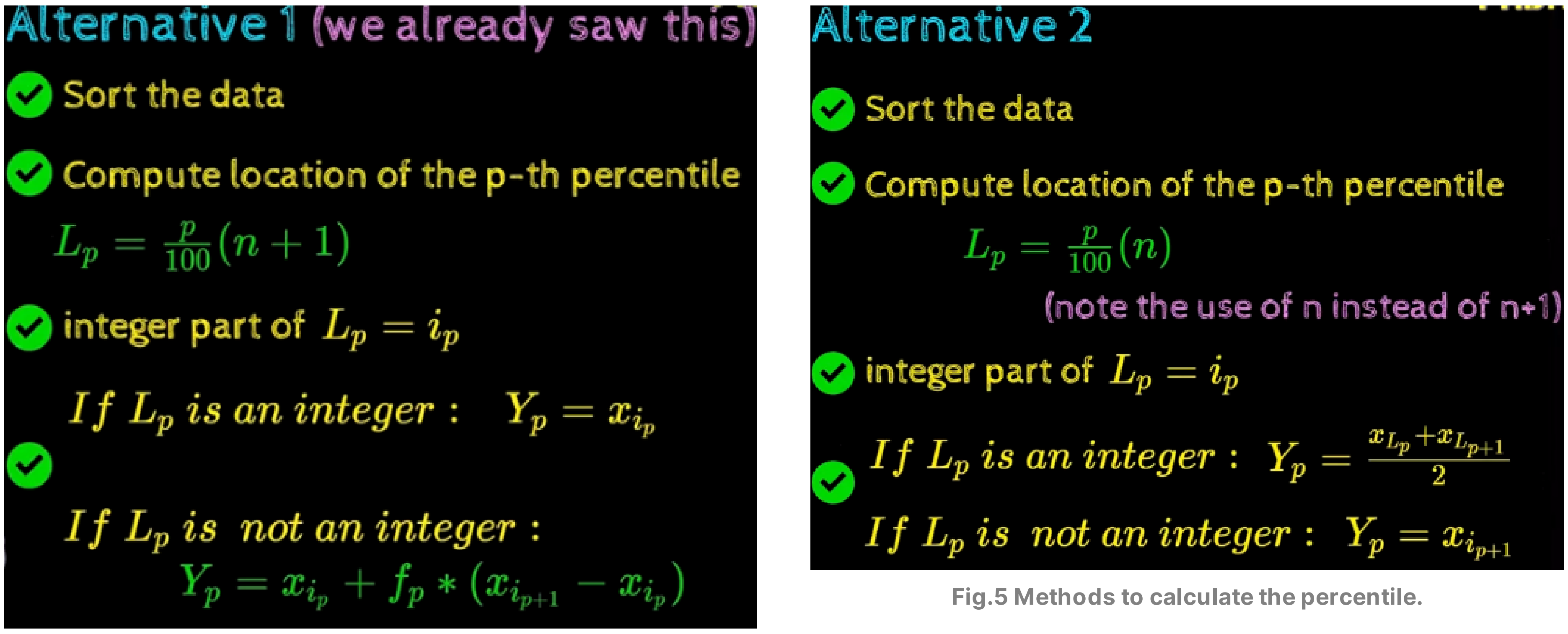
$i_p = 20, f_p = 0.8$

$Y_{80} = x_{20} + 0.8 * (x_{21} - x_{20}) = 61.6$

Fig.4 Example for percentile calculation.

Alternative methods of computing percentile - Part1

def (alternate2) percentile - (t = 7:02)



Alternative 1 (we already saw this)

✓ Sort the data

✓ Compute location of the p-th percentile

$L_p = \frac{p}{100}(n+1)$

✓ integer part of $L_p = i_p$

✓ $If\ L_p\ is\ an\ integer:\quad Y_p = x_{i_p}$

$If\ L_p\ is\ not\ an\ integer:$

$Y_p = x_{i_p} + f_p * (x_{i_{p+1}} - x_{i_p})$

Alternative 2

✓ Sort the data

✓ Compute location of the p-th percentile

$L_p = \frac{p}{100}(n)$

(note the use of n instead of n+1)

✓ integer part of $L_p = i_p$

✓ $If\ L_p\ is\ an\ integer:\quad Y_p = \frac{x_{L_p} + x_{L_{p+1}}}{2}$

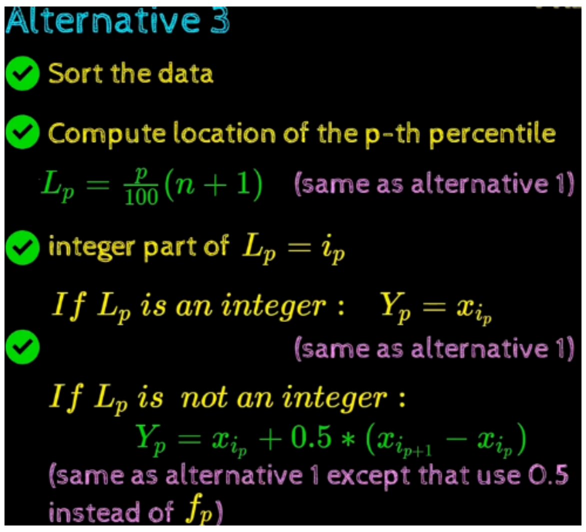
$If\ L_p\ is\ not\ an\ integer:\quad Y_p = x_{i_{p+1}}$

Fig.5 Methods to calculate the percentile.

In the second method, there is a slight change in the definition.

The pth percentile is the value in data such that p% of the values are less than or equal to it and at lead (100 - p)% of the values are greater than equal to it.

Alternative methods of computing percentile - Part2



Alternative 3

✓ Sort the data

✓ Compute location of the p-th percentile

$L_p = \frac{p}{100}(n+1)$ (same as alternative 1)

✓ integer part of $L_p = i_p$

✓ $If\ L_p\ is\ an\ integer: \quad Y_p = x_{i_p}$ (same as alternative 1)

$If\ L_p\ is\ not\ an\ integer:$

$Y_p = x_{i_p} + 0.5 * (x_{i_{p+1}} - x_{i_p})$

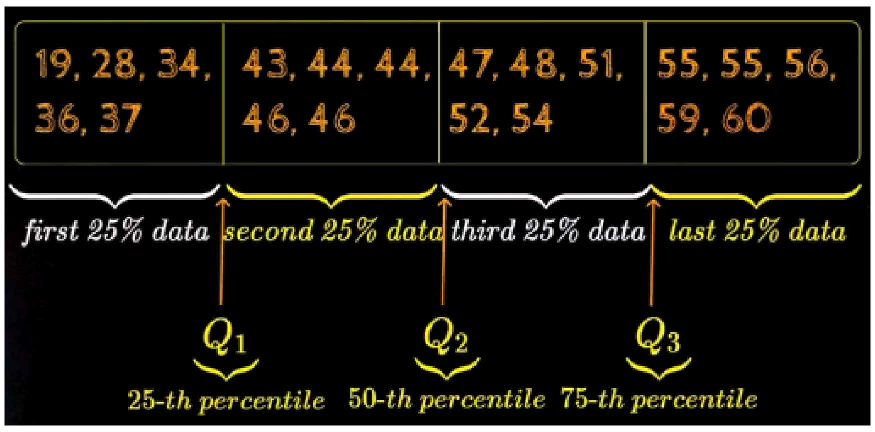
(same as alternative 1 except that use 0.5 instead of f_p)

Fig.6 Method3 to calculate the percentile.

The first method is more precise.

Frequently used percentile

Quartiles divide the data into 4 equal part, using Q1 ,Q2 ,Q3 .



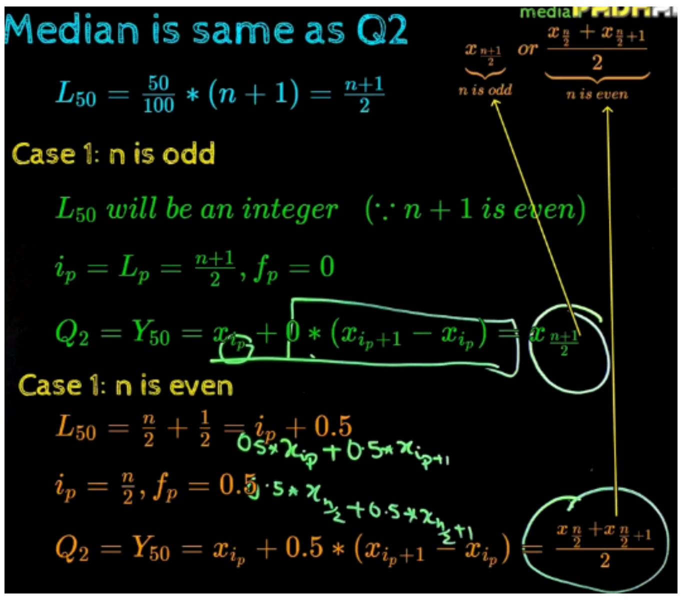
19, 28, 34, 36, 37	43, 44, 44, 46, 46	47, 48, 51, 52, 54	55, 55, 56, 59, 60

Fig.7 Quartiles.

Computation involves finding the 25th,50th and 75th percentiles as Q1 ,Q2 ,Q3 respectively.

Median is same as Q2

Proof:



Median is same as Q2

$\underbrace{x_{\frac{n+1}{2}}}_{n\ is\ odd}$ or $\underbrace{\frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}}_{n\ is\ even}$

$L_{50} = \frac{50}{100} * (n+1) = \frac{n+1}{2}$

Case 1: n is odd

$L_{50}\ will\ be\ an\ integer \quad (\because n+1\ is\ even)$

$i_p = L_p = \frac{n+1}{2}, f_p = 0$

$Q_2 = Y_{50} = x_{i_p} + 0 * (x_{i_p+1} - x_{i_p}) = x_{\frac{n+1}{2}}$

Case 1: n is even

$L_{50} = \frac{n}{2} + \frac{1}{2} = i_p + 0.5$

$0.5 * x_{i_p} + 0.5 * x_{i_{p+1}}$

$i_p = \frac{n}{2}, f_p = 0.5$

$0.5 * x_{\frac{n}{2}} + 0.5 * x_{\frac{n}{2}+1}$

$Q_2 = Y_{50} = x_{i_p} + 0.5 * (x_{i_p+1} - x_{i_p}) = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$

Fig.8 Proof of median = Q2

Quintiles divide the data into five equal parts.Calculation involves finding the 20th,40th,60th,80th percentiles.

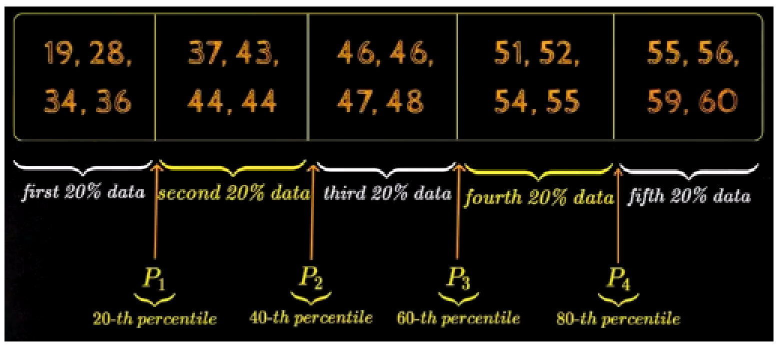


Fig.9 Quintiles.

Deciles divide the data into 10 equal parts. Calculation involves finding the 9 percentiles that divide the data into 10 equal parts.

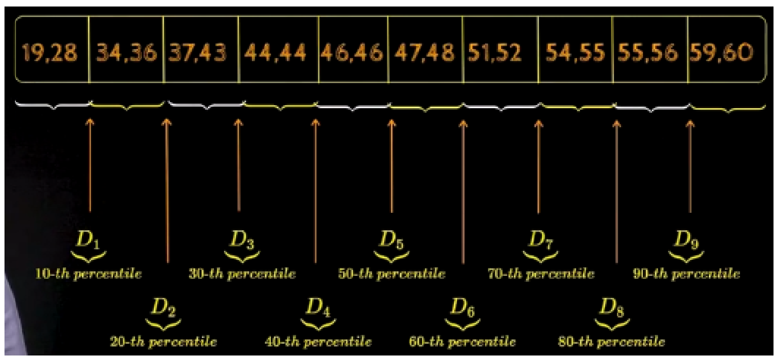
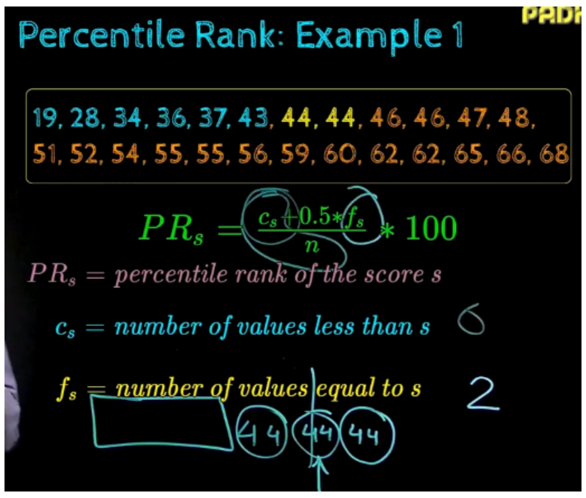


Fig.10 Deciles.

Compute the percentile rank of a value in the data

def percentile rank - (t = 0:40)

Percentile rank of a value is the percentage of data values that are less than or equal to it. Percentile rank of a score is computed by:



Percentile Rank: Example 1

19, 28, 34, 36, 37, 43, 44, 44, 46, 46, 47, 48,
51, 52, 54, 55, 55, 56, 59, 60, 62, 62, 65, 66, 68

$$PR_s = \frac{c_s + 0.5 * f_s}{n} * 100$$

PR_s = percentile rank of the score s

c_s = number of values less than s 6

f_s = number of values equal to s 2

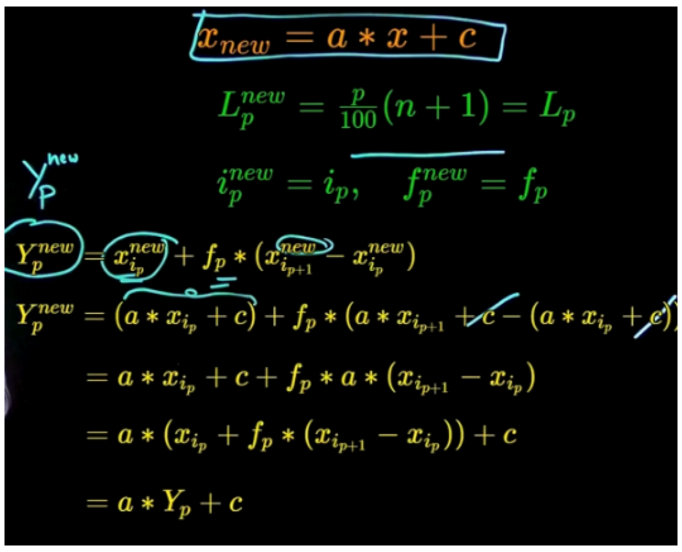
44 44 44

Fig.11 Formula to calculate percentile rank.

Here, PR_{44} = (6 + 0.5 * 2) / 25 = 28. This is interpreted as there are 28% values lesser than 44 and 72% values lie above 44. (note: c_s is the number of values strictly less than the score s). The percentile thus calculated is rounded off to the next whole number.

Effect of transformation on percentiles

Transformation involves scaling and shifting of data points.



$$\boxed{x_{new} = a * x + c}$$

$$L_p^{new} = \frac{p}{100}(n+1) = L_p$$

$$Y_p^{new}$$

$$i_p^{new} = i_p, \quad \overline{f_p^{new} = f_p}$$

$$Y_p^{new} = x_{i_p}^{new} + f_p * (x_{i_{p+1}}^{new} - x_{i_p}^{new})$$

$$Y_p^{new} = (a * x_{i_p} + c) + f_p * (a * x_{i_{p+1}} + \cancel{c} - (a * x_{i_p} + \cancel{c})$$

$$= a * x_{i_p} + c + f_p * a * (x_{i_{p+1}} - x_{i_p})$$

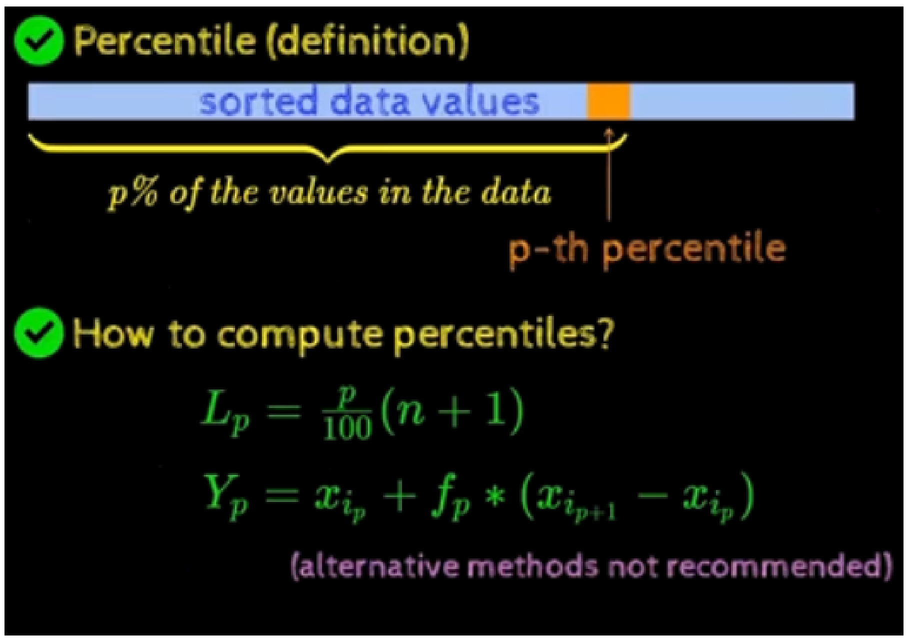
$$= a * (x_{i_p} + f_p * (x_{i_{p+1}} - x_{i_p})) + c$$

$$= a * Y_p + c$$

Fig.12 relationship between old and new percentile values.

The location of pth percentile does not change since the relative ordering of the elements does not change. The new value is calculated using the same formula used for percentile, with the difference of using the transformed elements. The percentile value is transformed by the same transformations applied on the data.

Summary - percentiles



✓ Percentile (definition)

sorted data values

$p\%$ *of the values in the data*

p-th percentile

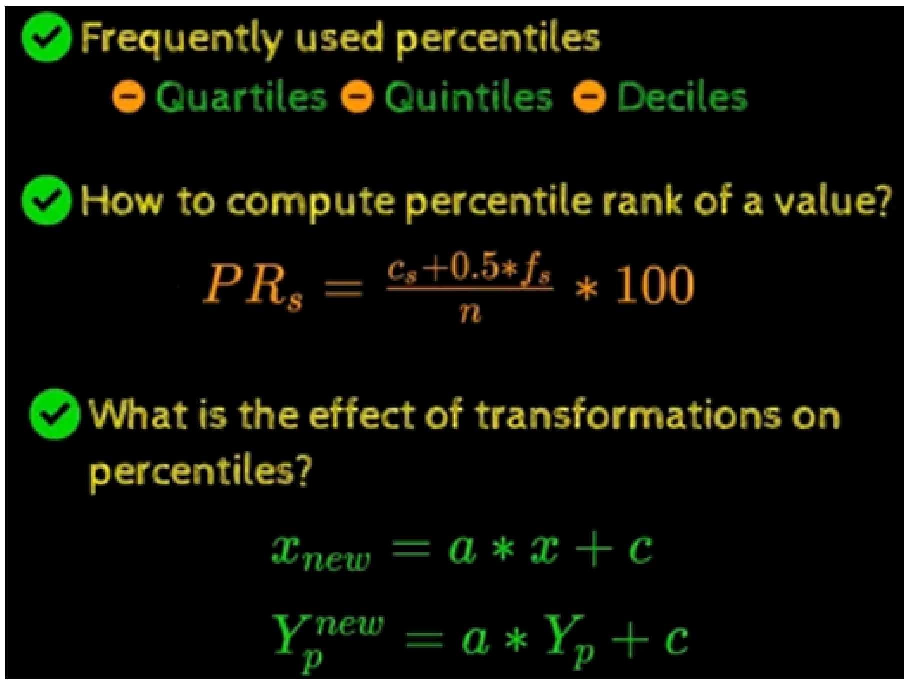
✓ How to compute percentiles?

$$L_p = \frac{p}{100}(n+1)$$

$$Y_p = x_{i_p} + f_p * (x_{i_{p+1}} - x_{i_p})$$

(alternative methods not recommended)

Fig.13 Percentiles summary



✓ Frequently used percentiles

● Quartiles ● Quintiles ● Deciles

✓ How to compute percentile rank of a value?

$$PR_s = \frac{c_s + 0.5 * f_s}{n} * 100$$

✓ What is the effect of transformations on percentiles?

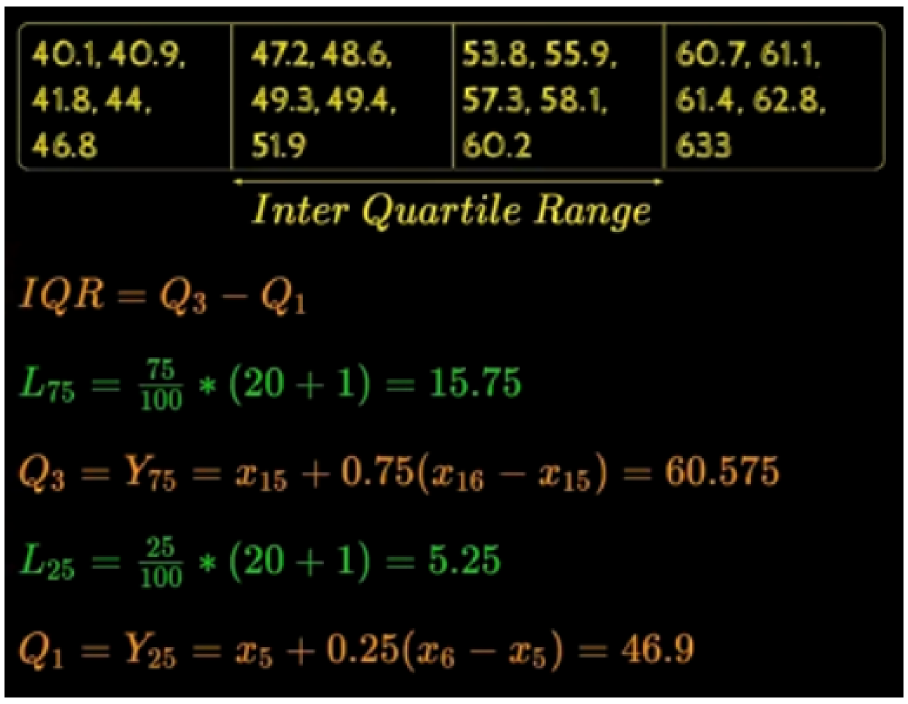
$$x_{new} = a * x + c$$

$$Y_p^{new} = a * Y_p + c$$

- The data point corresponding to p^{th} percentile of a data set is greater than or equal to p% of data points.
- First the position of the p^{th} percentile is calculated using $L_p = \frac{p}{100}(n+1)$. Then, the number representing this position is calculated using $Y_p = x_p + f_p * (x_{ip+1} - x_{ip})$.
- Quartiles, quintiles and deciles are the frequently used percentiles. The median is equal to the second quartile and the fifth decile.
- The percentage rank of a score represents the percentage of values that are lesser than or equal to the given score in the data set.
- When the data is scaled by ‘a’ and shifted by ‘c’, the new percentile is calculated by scaling and shifting the old percentile by a and c respectively.

Measures of spread

- Measures of centrality do not contain information about spread or variability in the data.
- **Ranges** is the difference between min and max values in the data. Higher the range, higher the spread in the data. Range is exaggerated in the presence of outliers. Thus, range is **sensitive to outliers.**
- **(Inter Quartile Range)IQR** is used to overcome the sensitivity of range to outliers. IQR is the difference between Q3 and Q1. It represents the range of 50% of the data. IQR is **not sensitive** to outliers.



40.1, 40.9, 41.8, 44, 46.8	47.2, 48.6, 49.3, 49.4, 51.9	53.8, 55.9, 57.3, 58.1, 60.2	60.7, 61.1, 61.4, 62.8, 633

Inter Quartile Range

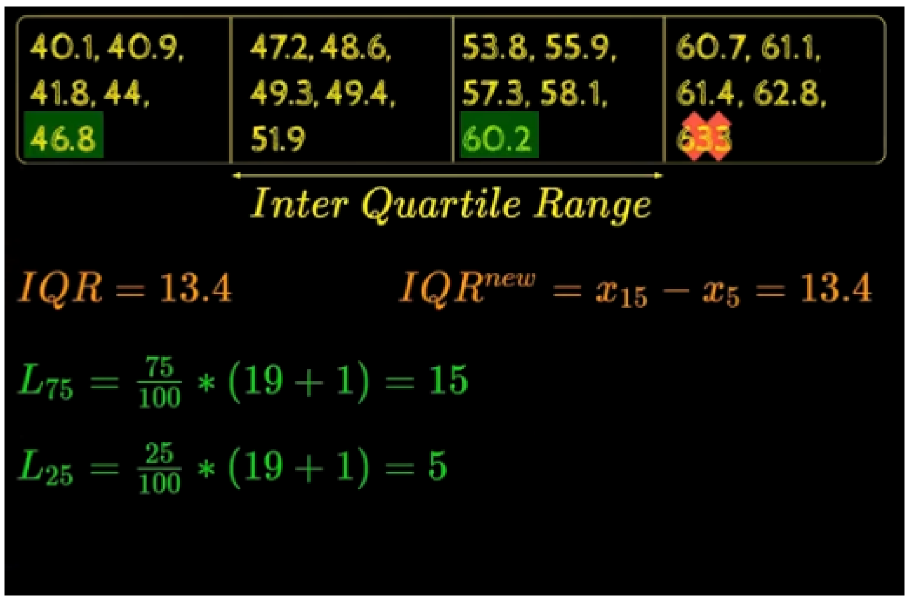
$$IQR = Q_3 - Q_1$$

$$L_{75} = \frac{75}{100} * (20+1) = 15.75$$

$$Q_3 = Y_{75} = x_{15} + 0.75(x_{16} - x_{15}) = 60.575$$

$$L_{25} = \frac{25}{100} * (20+1) = 5.25$$

$$Q_1 = Y_{25} = x_5 + 0.25(x_6 - x_5) = 46.9$$



40.1, 40.9, 41.8, 44, 46.8	47.2, 48.6, 49.3, 49.4, 51.9	53.8, 55.9, 57.3, 58.1, 60.2	60.7, 61.1, 61.4, 62.8, ~~633~~

Inter Quartile Range

$$IQR = 13.4 \qquad IQR^{new} = x_{15} - x_5 = 13.4$$

$$L_{75} = \frac{75}{100} * (19+1) = 15$$

$$L_{25} = \frac{25}{100} * (19+1) = 5$$

Fig.14 Example calculation of IQR before and after removal of outlier.

Measures of spread (variance)

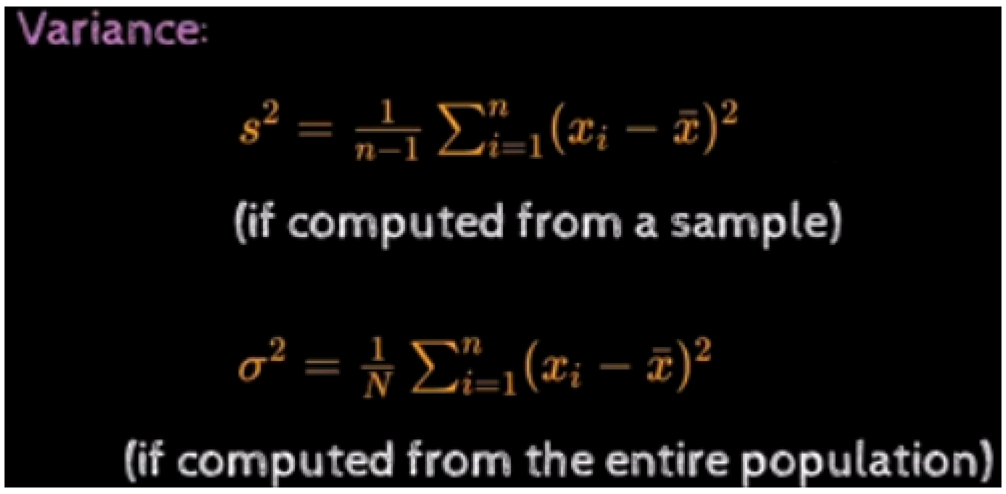
- Measures of spread represent how far are the values in the data from the typical value (mean) in the data. The convention is to use mean.
- **Deviation** from the mean can be calculated for all points and averaged. But, the sum of deviations from the mean is 0. Thus, the spread of the data cannot be inferred. This is because the positive deviations are cancelled out by the negative deviation.
- To resolve this **absolute deviation** measures can be used

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

or, **squared deviation** values can be used to and averaged over *n*.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

This is the preferred solution and is called the **variance.**



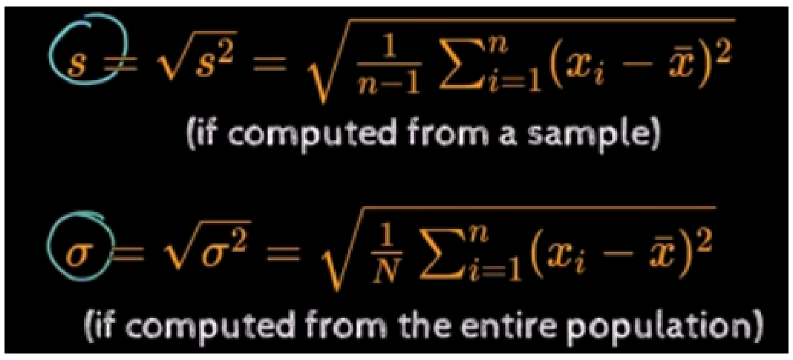
Variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(if computed from a sample)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(if computed from the entire population)



$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

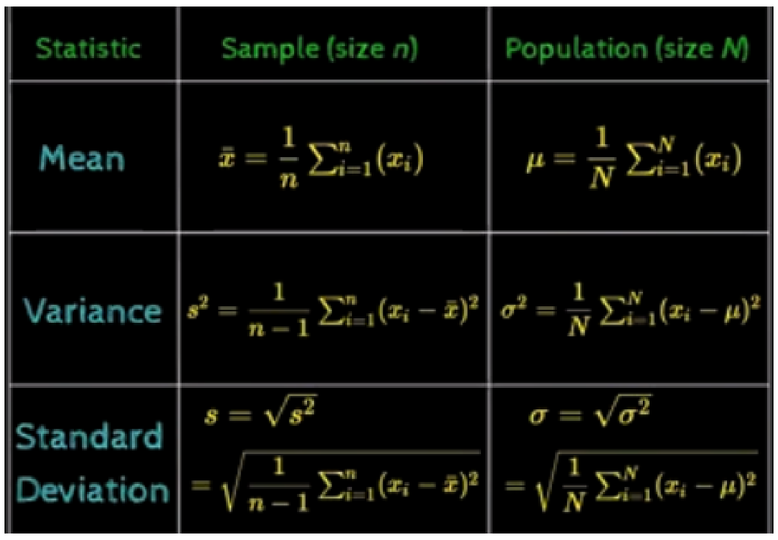
(if computed from a sample)

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(if computed from the entire population)

Fig.15 Computing variance and standard deviation (note the difference in denominator)

- By convention, s^2 is used for the sample and σ^2 for the population.
- **Observation :** variance is not measured in the same unit as the data. It is the squared unit. Therefore, **standard deviation** is used, it is square root of the variance and same unit as the data.



Statistic	Sample (size *n*)	Population (size *M*)
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i)$	$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$
Variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
Standard Deviation	$s = \sqrt{s^2}$ $= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$\sigma = \sqrt{\sigma^2}$ $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

Fig.16 Recap of notations.

Why do we square the deviations?

The square function has better properties than the abs function.

- It is a smooth curve and hence differentiable everywhere.
- Abs function is not differentiable at $x_i - \bar{x}$

In ML, functions that are differentiable are required.

The square function magnifies the contribution of outliers and suppresses the contribution of smaller values.

What does the variance tell us about the data?

Variance is considered as a **measure of consistency.** Smaller the value, better the consistency.

Effect of transformations on measures of spread

Range : it only gets scaled by 'a' and not shifted.

$$range_{new} = a * range$$

IQR : it is only scaled by 'a'.

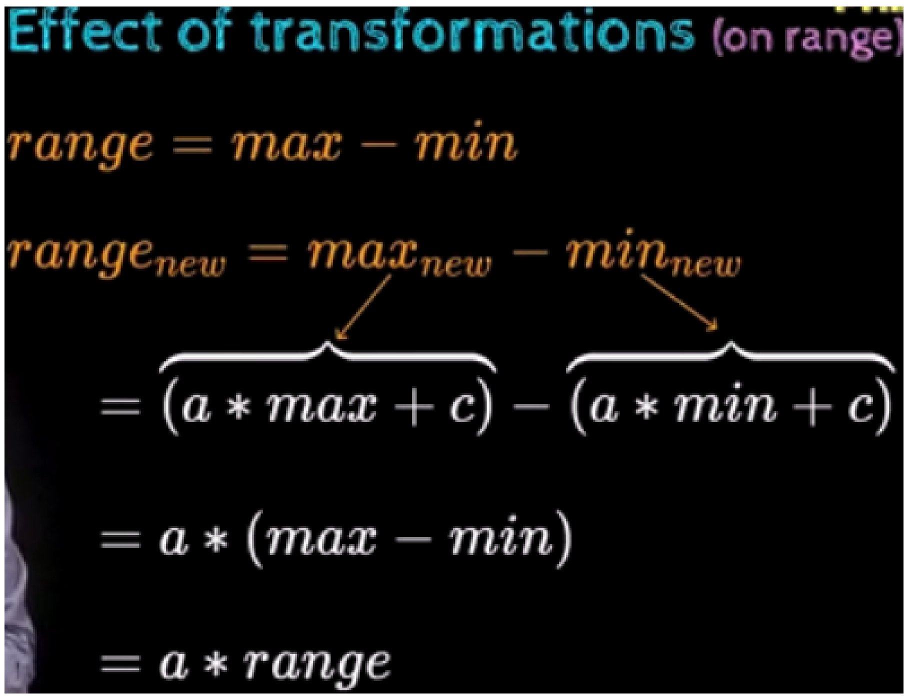
$$IQR_{new} = a * IQR$$

Variance : the new variance is scaled by a2

$$s_{new}^2 = a^2 * s^2$$

Standard deviation : it is scaled by a.

$$s_{new} = a * s$$



Effect of transformations (on range)

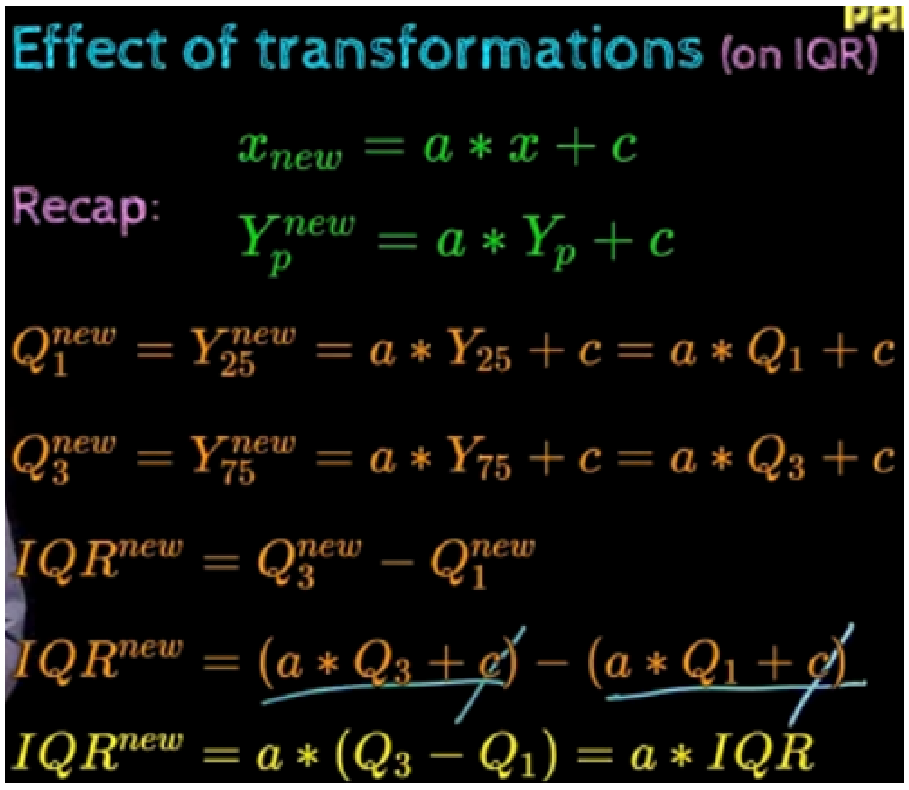
$$range = max - min$$

$$range_{new} = max_{new} - min_{new}$$

$$= \overbrace{(a * max + c)} - \overbrace{(a * min + c)}$$

$$= a * (max - min)$$

$$= a * range$$



Effect of transformations (on IQR)

Recap:
$$x_{new} = a * x + c$$
$$Y_p^{new} = a * Y_p + c$$

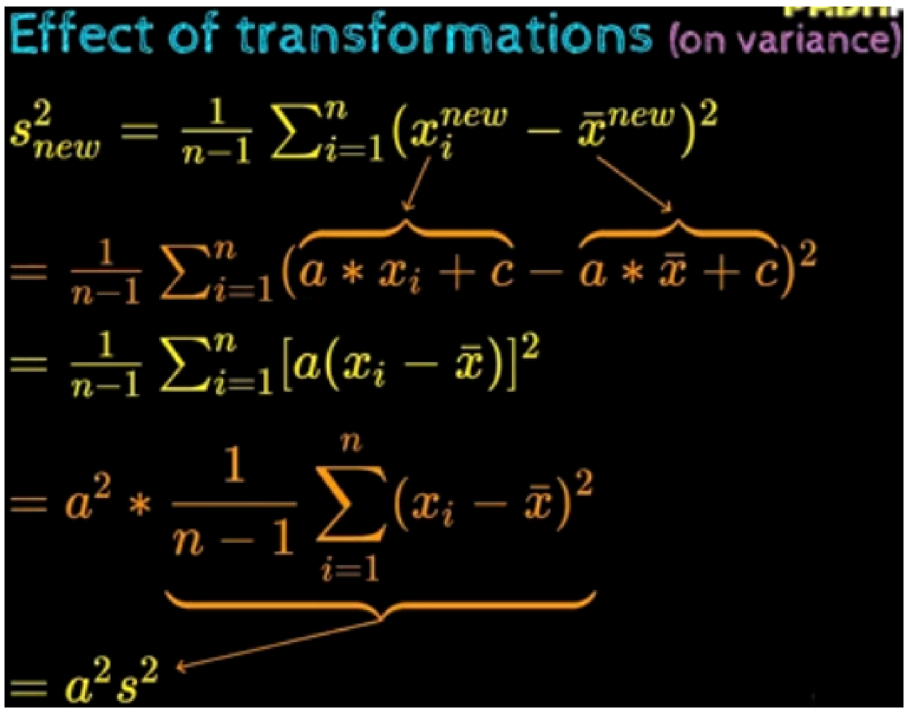
$$Q_1^{new} = Y_{25}^{new} = a * Y_{25} + c = a * Q_1 + c$$

$$Q_3^{new} = Y_{75}^{new} = a * Y_{75} + c = a * Q_3 + c$$

$$IQR^{new} = Q_3^{new} - Q_1^{new}$$

$$IQR^{new} = (a * Q_3 + \cancel{c}) - (a * Q_1 + \cancel{c})$$

$$IQR^{new} = a * (Q_3 - Q_1) = a * IQR$$



Effect of transformations (on variance)

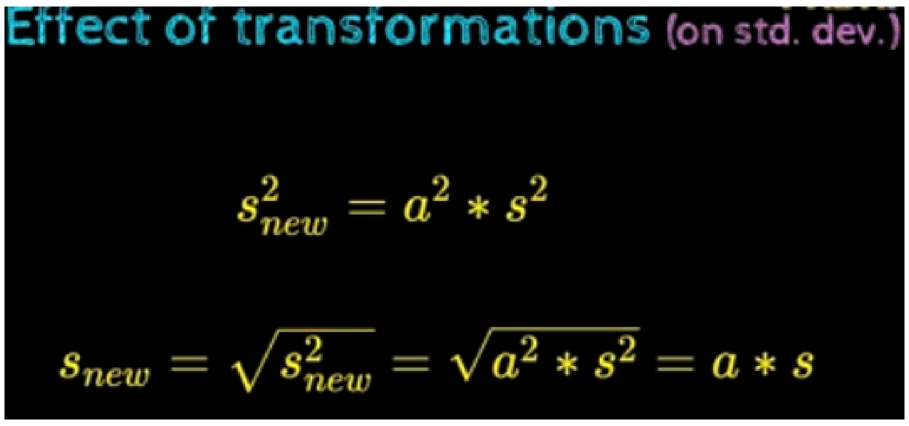
$$s^2_{new} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i^{new} - \bar{x}^{new})^2$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (\overbrace{a * x_i + c} - \overbrace{a * \bar{x} + c})^2$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} [a(x_i - \bar{x})]^2$$

$$= a^2 * \underbrace{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= a^2 s^2$$



Effect of transformations (on std. dev.)

$$s^2_{new} = a^2 * s^2$$

$$s_{new} = \sqrt{s^2_{new}} = \sqrt{a^2 * s^2} = a * s$$

Fig.17 Effects of transformations on measures of spread.

How do you use mean and variance to standardise data?

- **Standardising data :** the aim is to find how many standard deviations away from the mean is a given point. The distances are expressed in units of standard deviations instead of absolute values. Any point is the data can be expressed in terms of mean $\bar{x}$, standard deviation ‘s’ and z_i, the number of standard deviations.

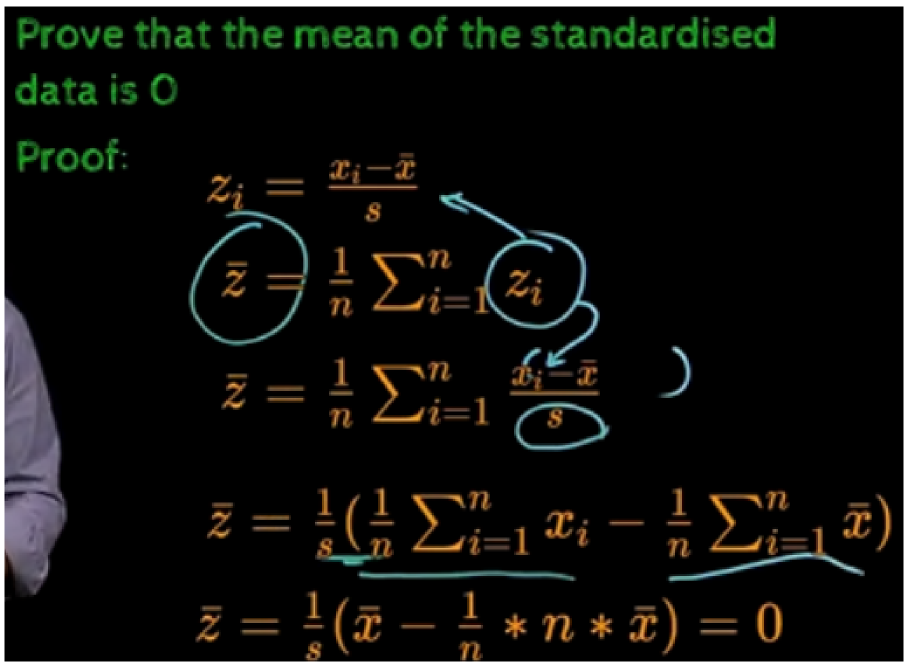
$$x_i = \bar{x} + z_i * s$$

where,

$$z_i = \frac{x_i - \bar{x}}{s}$$

- z is called the **z score** of a data point. It represents the distance of the point from the mean value and the direction. This is the standard form of the data.
- In ML systems, the ranges of inputs have to be standardised to avoid biases in training.

The mean of standardised data is 0 and standard deviation is 1.



Prove that the mean of the standardised data is 0

Proof:

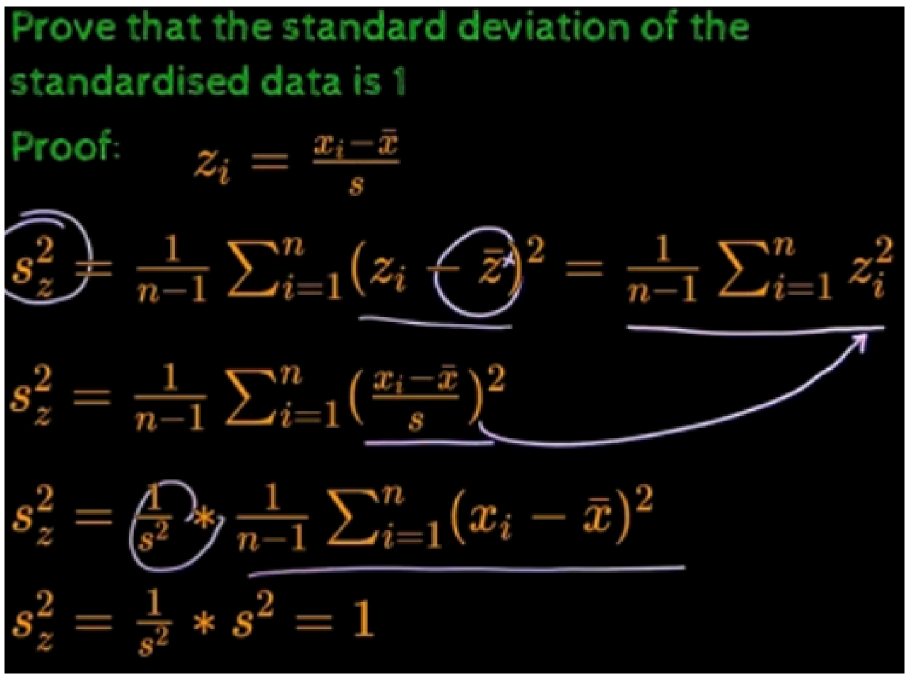
$$z_i = \frac{x_i-\bar{x}}{s}$$

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i-\bar{x}}{s}$$

$$\bar{z} = \frac{1}{s}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} \bar{x}\right)$$

$$\bar{z} = \frac{1}{s}\left(\bar{x} - \frac{1}{n} * n * \bar{x}\right) = 0$$



Prove that the standard deviation of the standardised data is 1

Proof: $z_i = \frac{x_i-\bar{x}}{s}$

$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{n-1}\sum_{i=1}^{n} z_i^2$$

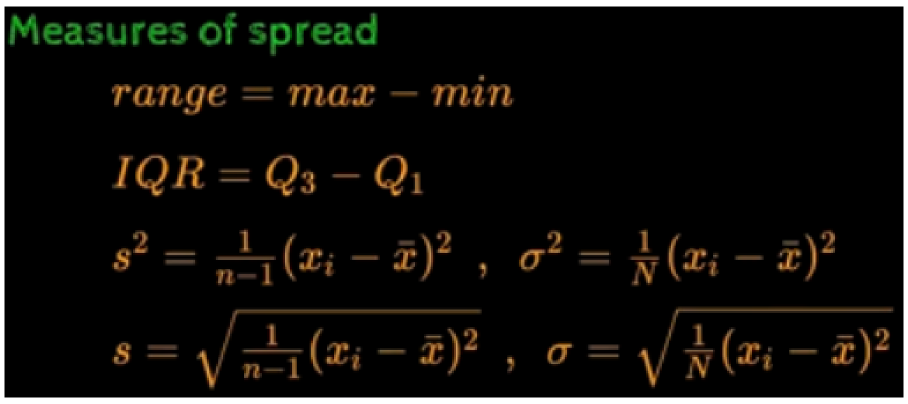
$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^2$$

$$s_z^2 = \frac{1}{s^2} * \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$s_z^2 = \frac{1}{s^2} * s^2 = 1$$

Fig.18 Proof for mean=0 and SD=1 for standardised data.

Summary - measures of spread



Measures of spread

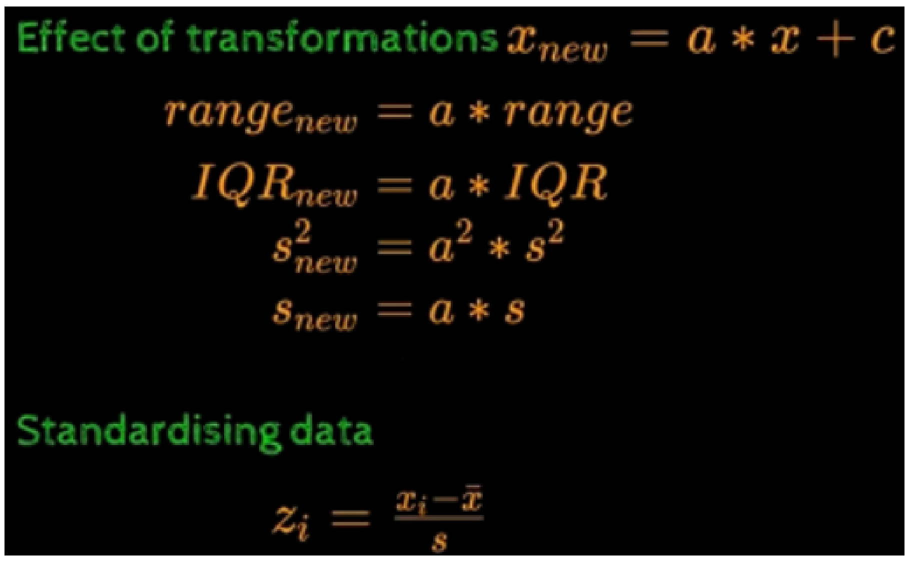
$$range = max - min$$

$$IQR = Q_3 - Q_1$$

$$s^2 = \frac{1}{n-1}(x_i-\bar{x})^2 \quad , \quad \sigma^2 = \frac{1}{N}(x_i-\bar{x})^2$$

$$s = \sqrt{\frac{1}{n-1}(x_i-\bar{x})^2} \quad , \quad \sigma = \sqrt{\frac{1}{N}(x_i-\bar{x})^2}$$

Fig.19 Formulae for measures of spread, effects of transformation and standardizing data.



Effect of transformations $x_{new} = a * x + c$

$$range_{new} = a * range$$

$$IQR_{new} = a * IQR$$

$$s^2_{new} = a^2 * s^2$$

$$s_{new} = a * s$$

Standardising data

$$z_i = \frac{x_i-\bar{x}}{s}$$

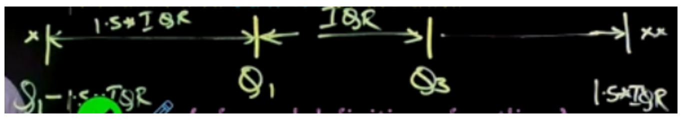
- Measures of spread represent the consistency in the data.
- Except IQR all measures are sensitive to outliers.
- Except variance, all measures have the same unit as the original data.
- Transformation of data results only in scaling of measures of spread by ‘a’ and variance is scaled up ‘a2’.
- After standardization, the data has zero mean and unit variance.

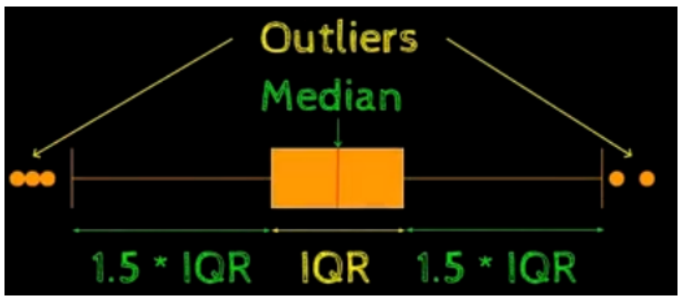
What are box plots?

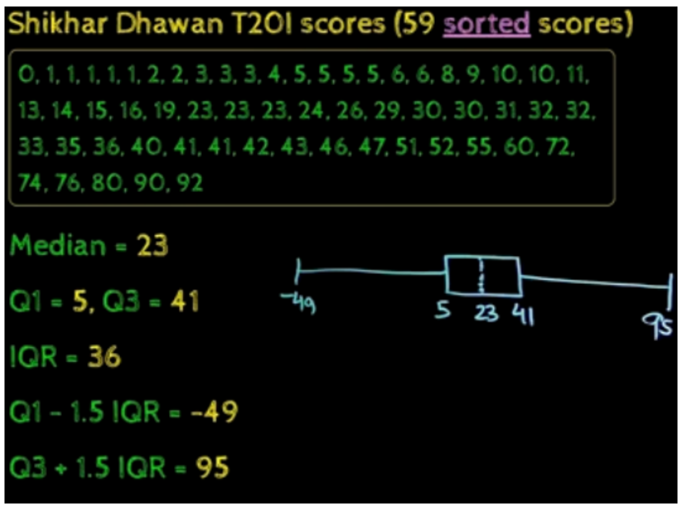
def outlier - (t = 0:55)

- Box plots are used for visualising spread, median and outliers in the data. A point x is said to be an outlier if,

$$x < Q_1 - 1.5 * IQR (or) x > Q_3 + 1.5 * IQR$$







Shikhar Dhawan T20I scores (59 sorted scores)

0, 1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 4, 5, 5, 5, 5, 6, 6, 8, 9, 10, 10, 11, 13, 14, 15, 16, 19, 23, 23, 23, 24, 26, 29, 30, 30, 31, 32, 32, 33, 35, 36, 40, 41, 41, 42, 43, 46, 47, 51, 52, 55, 60, 72, 74, 76, 80, 90, 92

Median = 23

Q1 = 5, Q3 = 41

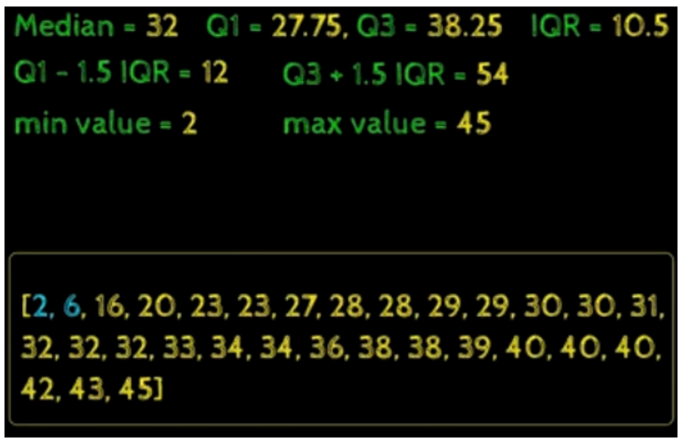
IQR = 36

Q1 - 1.5 IQR = -49

Q3 + 1.5 IQR = 95

Fig.20 a. Understanding outliers ,b. Box plots, c. example of box plot

- 50% of the data lies in the IQR, any point very far from this range would thus be an outlier.
- In a box plot the markers of 1.5*IQR are called **whiskers.** Points that lie beyond the whiskers are the outliers.
- Consider the data points and the measures required to draw a box plot:



Median = 32 Q1 = 27.75, Q3 = 38.25 IQR = 10.5

Q1 - 1.5 IQR = 12 Q3 + 1.5 IQR = 54

min value = 2 max value = 45

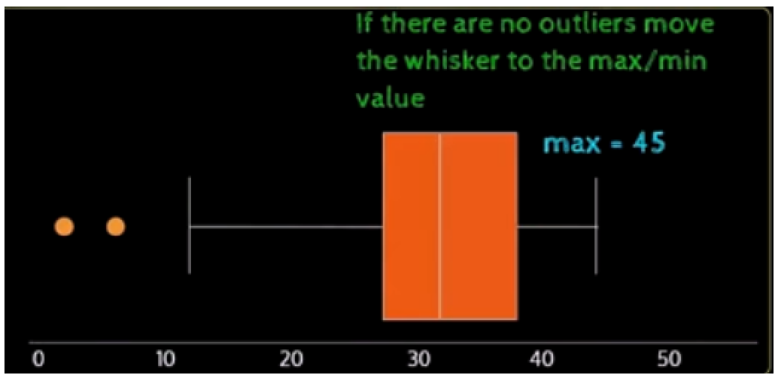
[2, 6, 16, 20, 23, 23, 27, 28, 28, 29, 29, 30, 30, 31, 32, 32, 32, 33, 34, 34, 36, 38, 38, 39, 40, 40, 40, 42, 43, 45]

Fig.21 An example dataset and measurements.

Here, 2 and 6 are outliers.

Box Plots (Variant 1):

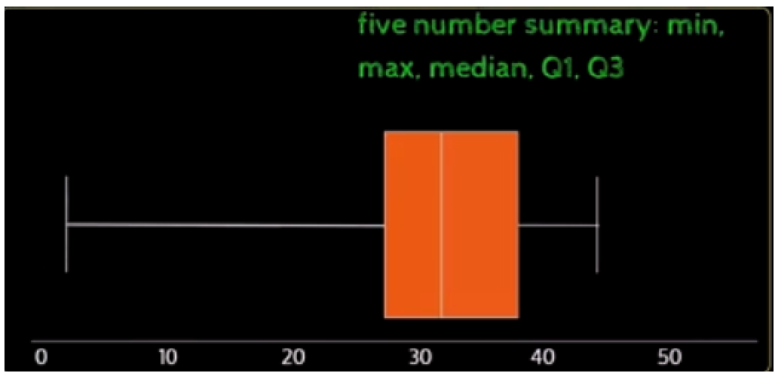
- If there are no outliers, instead of placing the whiskers at the theoretical outlier boundary, they are placed at the min and max values.



If there are no outliers move the whisker to the max/min value

Fig.22 Varient1 of box plot.

Box plots (Variant 2): five number summary of the data



five number summary: min, max, median, Q1, Q3

Fig.23 Box plot with five number summary of the data.

- The whiskers are placed at min and max values.

- The drawback here is that it does not show the outliers.

Box Plots (Variant 3): min and max values are calculated excluding the outliers

- Here, min value = 16 is used instead of 2.

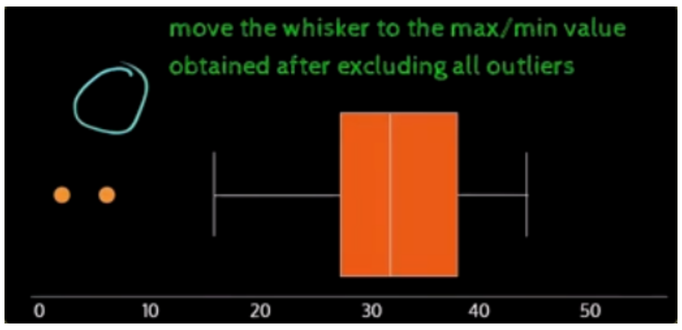


Fig.24 Variant 3

Box Plots : skewness in the data

- In right skewed data, there is a long tail in the RHS and the median lies towards the LHS.
- This can be inferred from the box plot as follows:

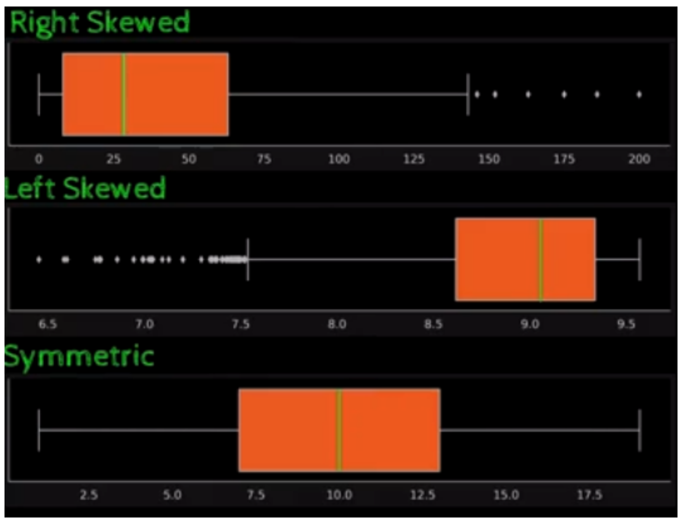
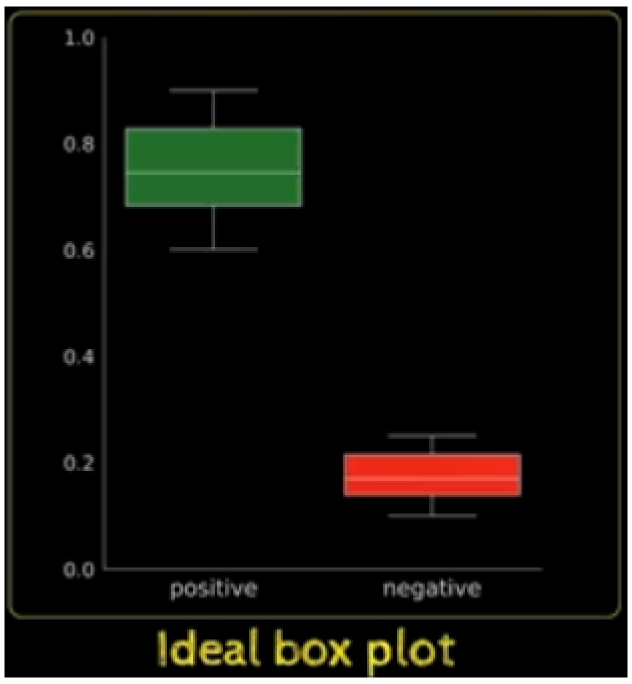


Fig.25 Skewness in data.

- The median lies towards the LHS of the box and a lot of outliers are found towards the RHS.
- In left skewed data the median lies towards the RHS of the box, the long tail is on the LHS and a lot of outliers are found on this side.
- The median would be in the center of the box in perfectly symmetric data.

Box Plots (usage in ML)

- For classification ML problems, the predictions can be visualized using box plots.
- Box plots can be used to visually compare the performance of ML systems.
- Ideal case is to have them as far as possible. There will not be any overlap in IQR of scores assigned to the classes.



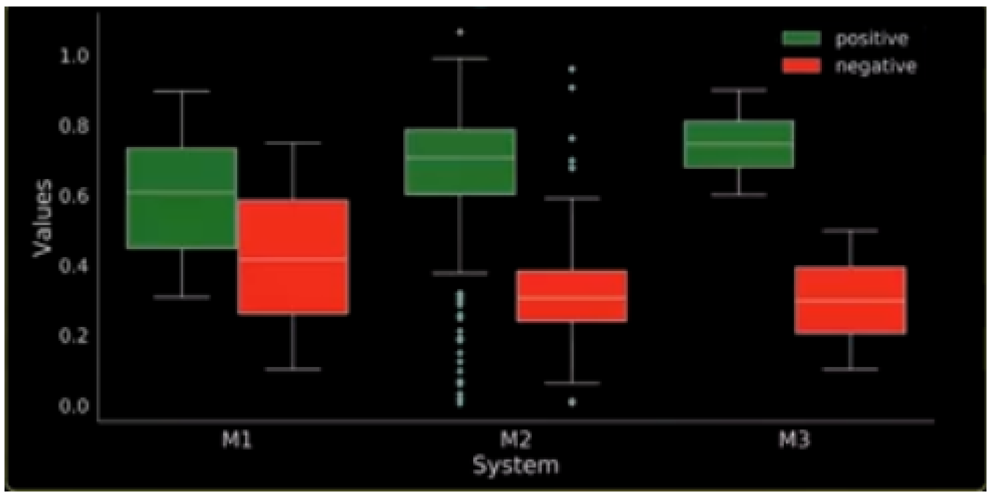


Fig.26 a. Ideal box plot, b. Comparison of different ML systems.

The above box plots are for classification of positive and negative reviews.

- The first model M1 (fig.26 b) is bad as the IQRs of the two box plots overlap. Here both positive and negative reviews are being assigned scores near 0.5.
- In the second model M2, the IQRs do not overlap, but there are outliers in each classification. i.e. there are negative reviews being assigned high scores and positive reviews being assigned low scores, implying, the model is not clearly differentiating between positive and negative reviews.
- M3 is very close to the ideal system. The model assigns very high scores to the positive reviews, low scores to the negative reviews and there are no outliers.

Thus, box plots can be used to compare ML systems.

Summary Box Plots

- The median, IQ, outliers and data skewness can be depicted in a box plot.
- Q2 corresponds to the median of the data. The points beyond the whiskers are the outliers.
- Box plots can be used to compare performance of different ML classification systems.

MCQ : percentile

1. In right-skewed distributions
    1. the distance from Q1 to Q2 is larger than the distance from Q2 to Q3.
    2. the mean is smaller than the median.
    3. **the distance from Q1 to Q2 is smaller than the distance from Q2 to Q3.**
    4. the mode is larger than the mean.
2. Which of the following statements is incorrect with respect to median?
    1. It is equal to the mode in bell-shaped, symmetrical distributions.
    2. It is equal to the Second Quartile.
    3. It is a measure of central tendency.
    4. **The median is more affected by extreme values than the mean.**
3. Which of the following statistics always corresponds to the 75th percentile in a distribution?
    1. Median of Q2 and values lesser than Q2..
    2. **Third Quartile.**
    3. **Median of Q2 and values greater than Q2.**
    4. Median.
4. In a factory, the weight of the concrete poured into a mold by a machine follows a normal distribution with a mean of 1150 pounds and a standard deviation of 22 pounds. What weight falls at the 50th percentile?
    1. **1150**
    2. 1106
    3. 1128
    4. 1172
5. Suppose a student scores 90th percentile in a college entrance examination. This means
    1. The student correctly answered 90% of the questions.
    2. **The student scores better than 90% of the students in the exam.**
    3. The score of 90 is the highest score in the exam.
    4. **10% of the students scored better than the given student.**

6. In a project review of 1205 students, if a student finds that 47 projects received better scores than his, what percentile does the student come closest in the overall project score ranking?
    1. 90th percentile
    2. 3rd percentile
    3. **96th percentile**
    4. Cannot be determined from the given information.

MCQ : Measures of spread

1. Median and IQR are used as measures of spread instead of mean and standard deviation when
    1. **The data has outliers.**
    2. **The data is skewed.**
    3. They can be used interchangeably, it is independent of data.
    4. None of the above
2. The match scores of players is summarised, select the most consistent player
    1. Player A with mean score of 80 and standard deviation of 40 runs
    2. **Player B with mean of 40 and standard deviation of 10 runs.**
    3. Player C with mean of 70 and standard deviation of 30 runs.
    4. None of the above.
3. Standard deviation cannot be zero
    1. True
    2. **False**
4. On addition of a large outlier
    1. Mean, IQR and standard deviation decrease
    2. **Mean, range and standard deviation increase**
    3. **Median and IQR stay relatively unchanged**
    4. Mean, median and mode reduce.
5. The average number of days a student is absent is 19 with a standard deviation of 4 days. How many standard deviations away from the mean is the student who is absent for three consecutive days?
    1. **-4**
    2. 4
    3. 2 standard deviations
    4. Three days is less than the standard deviation of four days, therefore, one standard deviation.